
Training Auto-encoders Effectively via Eliminating Task-irrelevant Input Variables

Hui Shen*, Dehua Li and Hong Wu

School of Automation
Huazhong University of Science & Technology
1037 Luoyu Road, Wuhan, Hubei, 430074 China
e-mail: shenhui@hust.edu.cn *Corresponding author;
lidehua1946@sina.com;
wolfe-wu@hust.edu.cn

Zhaoxiang Zang*

Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering
China Three Gorges University
Yichang Hubei, 443002, China
e-mail: zxzang@gmail.com *Corresponding author

Abstract: Auto-encoders are often used as building blocks of deep network classifier to learn feature extractors, but task-irrelevant information in the input data may lead to bad extractors and result in poor generalization performance of the network. In this paper, via dropping the task-irrelevant input variables the performance of auto-encoders can be obviously improved. Specifically, an importance-based variable selection method is proposed to aim at finding the task-irrelevant input variables and dropping them. It firstly estimates importance of each variable, and then drops the variables with importance value lower than a threshold. In order to obtain better performance, the method can be employed for each layer of stacked auto-encoders. Experimental results show that when combined with our method the stacked denoising auto-encoders achieves significantly improved performance on three challenging datasets.

Keywords: feature learning; deep learning; neural network; auto-encoder; stacked auto-encoders; variable selection; feature selection; unsupervised training

Biographical notes: Hui Shen is currently a PhD student in School of Automation, Huazhong University of Science and Technology, China. She obtained her bachelor's in Wuhan Polytechnic University, China. And got her master's in Huazhong University of Science and Technology, China. Her main fields of interest are neural network, deep learning and machine learning.

Dehua Li is a professor in School of Automation, Huazhong University of Science and Technology, China. He got his bachelor's from Wuhan University in 1970. And he spent one year as senior visiting scholar in AI department of University of Edinburgh, UK. His research interests including AI, neotic science, and machine learning.

Hong Wu obtained his bachelor's in computer science and technology from Wuhan University, China. And got his master's in Huazhong University of Science and Technology, China. Now he is a PhD student of the School of Automation, Huazhong University of Science and Technology, China. His main research fields are neural network, deep learning and machine learning.

Zhaoxiang Zang is an associate professor in College of Computer and Information Technology, China Three Gorges University, Yichang Hubei, China, and he is member of Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, China. He obtained his master's and PhD from Huazhong University of Science and Technology, China. His research interests mainly in the fields of machine learning, computational intelligence and computer game intelligence.

1 Introduction

Neural networks are widely applied in various fields, such as oil exploration in Liu et al. (2011), speech

recognition in Arisoy et al. (2015), temperature control in Admuthe and Chile (2015), and so on. As a kind of neural network, deep neural network has become an increasingly popular research field. And training auto-

encoders to learn useful feature extractors to initialize a deep neural network is a widely used approach. An auto-encoder is comprised by an encoder and a decoder. Given an input example, the encoder, which consists of a group of feature extractors, produces features that constitute an abstract representation or code of the example, while the decoder reconstructs the example from the code. Training an auto-encoder is to minimize the difference between the input example and its reconstruction. Details of auto-encoders and their applications in deep learning can be found in Hinton et al. (2006); Bengio (2013); Ciresan et al. (2012); Bengio et al. (2013). Auto-encoders are usually implemented with neural networks, but over-complete (higher dimensional hidden layer than the input layer) and unconstrained auto-encoders may learn identical mapping, which results in useless features. Therefore, many regularized auto-encoders were proposed to learn good feature extractors, such as sparse auto-encoders (Lee et al., 2008; Boureau et al., 2008; Ng, 2011), contractive auto-encoders (CAEs) (Rifai et al., 2011a,b,c), denoising auto-encoders (DAEs) (Vincent et al., 2008, 2010), marginalized denoising auto-encoders (mDAEs) (Chen et al., 2012, 2014), and so on.

Because of the unsupervised training, auto-encoders attempts to capture everything in input data, including task-irrelevant information if there exists (Vincent et al., 2010). However, learning task-irrelevant information may waste the computation resources and the capability of networks, and is easy to cause overfitting. Therefore, eliminating task-irrelevant information becomes one of the ways to obtain better performance and reduce computation cost. Bengio et al. explored supervised pre-training (Bengio et al., 2007), and they concluded that partially supervised pre-training (alternately perform supervised and unsupervised training of an auto-encoder) can lead auto-encoders to learn better feature extractors when much task-irrelevant information is contained in training data. Shon et al. proposed point-wise gated Boltzmann machines (PGBM) (Sohn et al., 2013). They used a hidden layer containing two-group units to model the foreground and background respectively, then took the foreground group to extract task-relevant features. PGMB achieves the state-of-the-art performance on several benchmark datasets. Inspired by human attention, Wang et al. proposed the attentional neural network (aNN) (Wang et al., 2014). They used a segmentation module to iteratively segment foreground from noisy input via a feedback loop, and employed a deep network on the foreground for classification. aNN also achieves the state-of-the-art performance on one benchmark dataset.

In this work, we try to drop the task-irrelevant variables by performing variable selection on input of auto-encoders and then execute unsupervised learning on the remaining variables. We introduce an importance-based variable selection method that evaluate importance of each variable and drop the variables with low importance. For obtaining better classification performance, the method is employed for

each layer of stacked auto-encoders, not only on the raw input for the first layer, which is different from methods mentioned above. The experiment results show that it helps stacked DAE (SDAE) achieve significantly improved performance on three challenging benchmark datasets.

The rest of this paper is organized as following. Preliminaries about DAE and SDAE are provided in section 2. The proposed variable selection method is described in detail in section 3. Experiments and results are reported in section 4, and conclusions are made in section 5. We will use following notations through out the paper. χ is a training dataset. Each element of χ contains a input example \mathbf{x} and a target label r such that $\mathbf{x} \in [0, 1]^M$ (or \mathbb{R}^M) and $r \in \{1, \dots, K\}$. For a vector \mathbf{x} , x_d is its d -th component. For the sake of simplicity, we use χ to denote the training dataset for an auto-encoder, no matter where the auto-encoder is located in a stacked auto-encoders.

2 Preliminaries

Auto-encoders(AEs) are often employed as building blocks of deep networks. An AE is a neural network that composed of three layers, an input layer, a hidden layer and an output layer. It tries to recover the input data \mathbf{x} from the hidden representation \mathbf{h} at the output layer. The motivation is that, if the input data can be reconstructed well enough, then it can be said that the hidden feature is a discription of the input data. An AE is also can be seen as a network that consists of an encoder and a decoder. Processing from input layer to hidden layer can be seen as an encoder, while from hidden layer to output layer a decoder.

During training, it firstly maps the the input data \mathbf{x} to hidden representation \mathbf{h} by the encoder:

$$\mathbf{h}_q = f_q(\mathbf{x}) = s_f(\mathbf{w}_q^T \mathbf{x} + b_q),$$

where \mathbf{w}_q^T is a weight vector, b_q is a bias and s_f is the activation function of the encoder, typically the sigmoid functions $s_f(z) = 1/(1 + e^{-z})$. Then reconstruct input y_d from the hidden representation \mathbf{h} through the decoder:

$$y_d = g_d(\mathbf{h}) = s_g(\mathbf{w}_d^T \mathbf{h} + c_d),$$

where c_d is a bias and s_g is the activation function of the decoder, which can be sigmoid function for binary input or an identity for continuous input.

Explicitly, the reconstruction \mathbf{y} should approach to original input \mathbf{x} as much as possible, this can be measured by the reconstruction error $L(\mathbf{x}, \mathbf{y})$ which is typically computed via squared error or cross-entropy. Which one to be chosen depends on the activation function of the decoder.

If s_g is an identity, i.e. $s_g(z) = z$, then

$$L(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2.$$

If s_g is simoid function, i.e. $s_g(z) = 1/(1 + e^{-z})$, then

$$L(\mathbf{x}, \mathbf{y}) = - \sum_{d=1}^{d=M} x_d \log(y_d) + (1 - x_d) \log(1 - y_d), (1)$$

where y_d is depend on the model parameters w_q, w_d, b_q and c_d when input data x is given.

Then the auto-encoder is trained to learn model parameters that minimize the reconstruction error $L(\mathbf{x}, \mathbf{y})$. Gradient descent can be utilized to optimized the error function during training. When the training is completed, the output layer with the weights of hidden to output are dropped and the learned representation feature holds in the hidden layer, which can be used for classification or used as the input of an other autoencoder to learn more abstract feature.

However, an AE may learn identity, which leads to obtain no useful featrues. In order to prevent this situation, AEs often utilize the configuration called ‘‘bottleneck’’ of which the quantity of hidden units lower than input units. An other approach is adding regular terms on the objective function to constrain the weights. Otherwise, using disturbed input data for training is also an effective means, like denoising autoencoder.

Denoising auto-encoder(DAE) is a variant of standard auto-encoder. It attempts to reconstruct the input \mathbf{x} from the encoded representation \mathbf{h} of noisy input $\tilde{\mathbf{x}}$ via a decoder. By disturbing the input \mathbf{x} , denoising auto-encoder tries to learn robust features that can successfully recover the perturbed values to reconstruct the original input data. If a DAE can recover the original input data from the code of corrupted input data, it can be said that the DAE has learned robust and stable features.

Stacked denoising auto-encoder(SDAE) is formed by stacking multiple single-layer DAEs for learning more abstract representations. SDAE can be used to effectively pre-train deep networks. In the process of pre-training by SDAE, the hidden features learned by lower-layer DAE are used as inputs for training next (upper-layer) DAE, and the encoders of DAEs are used to initialize weights in the deep network. See Vincent et al. (2010) for details.

3 Importance-based Variable Selection

According to equation (1), AEs belong to unsupervised learning without considering the label information. The hidden representation of an autoecoder is a description of the whole input data. AEs do not identify useful or unuseful information for classification. they attempt to capture all the information of the input, not only task-relevant information, but also task-irrelevant information if there exists. However, learning task-irrelevant information may waste computation resources and even cause over-fitting. Therefore, task-irrelevant information contained in input data should be reduced or eliminated for obtaining better performance.

To address this issue, an importance-based variable selection method is proposed to find the task-irrelevant variables and drop them. Briefly, the method is to evaluate the importance of each variable to classification and drop the variables with importance lower than a threshold. We exploit the sensitivity of the discriminant hyperplane to a variable to evaluate the importance of the variable. we argue that, the variables with higher sensitivity are more important for classification, these variables also possess higher importance value and should be reserved, while those variables with low importance(lower than a threshold) should be dropped. The details are described as follows.

We employ a trained Multinomial Logistic Regression(MLR) model as a pre-classifier to help us determine the importance of each input variable to classification. Multinomial Logistic Regression (MLR) is a simple log-linear classifier, and can be easily analyzed. Given an example \mathbf{x} , the MLR computes the posterior probability of each hypothesis via a softmax function, and takes the one with biggest posterior probability as prediction. see Bishop (2006); Hosmer Jr et al. (2013) for MLR in detail. We briefly introduce the softmax function, from which the importance notation can be deduced.

The softmax function can be written as

$$\sigma_i(\mathbf{x}) = \exp(\mathbf{w}_i^T \mathbf{x} + b_i) / \left[\sum_{c=1}^K \exp(\mathbf{w}_c^T \mathbf{x} + b_c) \right]$$

where \mathbf{w}_i and b_i are parameters, and $i \in \{1, \dots, K\}$ is the index of class. The predicted class of \mathbf{x} is obtained by $y = \arg \max_i \sigma_i(\mathbf{x})$.

The softmax function computes the estimated probability of the class label for a given input \mathbf{x} . Now we consider any two classes of class i and class j . We suppose that, the discriminant hyperplane between class i and class j is consisted of the points that with equivalent estimated probabilities of the two labels. Let

$$\frac{\exp(\mathbf{w}_i^T \mathbf{x} + b_i)}{\sum_{c=1}^K \exp(\mathbf{w}_c^T \mathbf{x} + b_c)} = \frac{\exp(\mathbf{w}_j^T \mathbf{x} + b_j)}{\sum_{c=1}^K \exp(\mathbf{w}_c^T \mathbf{x} + b_c)},$$

then we obtain the discriminant hyperplane:

$$(\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{x} + (b_i - b_j) = 0$$

After normalization, discriminant function between class i and class j can be written as

$$f_{i,j}(\mathbf{x}) = [(\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{x} + (b_i - b_j)] / \|\mathbf{w}_i - \mathbf{w}_j\|_2$$

In other words, all \mathbf{x} that satisfy $f_{i,j}(\mathbf{x}) = 0$ form a discriminant hyperplane between class i and class j . We denote the discriminant hyperplane as $\mathcal{H}_{i,j}$. The unit normal vector of $\mathcal{H}_{i,j}$ can be written as

$$\mathbf{v}_{i,j} = \frac{\mathbf{w}_i - \mathbf{w}_j}{\|\mathbf{w}_i - \mathbf{w}_j\|_2}. \quad (2)$$

Define the *sensitivity* of $f_{i,j}$ to an input variable as $|\frac{\partial f_{i,j}}{\partial x_d}|$, which reflects the influence of the variable to $f_{i,j}$.

Algorithm 1 Importance-based Variables Selection

Input: training dataset $\chi = \{\mathbf{x}^{(t)}, \mathbf{r}^{(t)}\}_{t=1}^N$, importance threshold c_{th}
Output: variables mask α
1: $\alpha \leftarrow (\mathbf{1}(\text{true}))_{d=1}^M$
2: **loop**
3: Train a new pre-classifier MLR with the masked training data $\{\alpha \odot \mathbf{x}^{(t)}, \mathbf{r}^{(t)}\}_{t=1}^N$.
4: **if** stop criterion **then**
5: **return** α
6: **else**
7: Update α according to (2), (3), (4), and (5).
8: **end if**
9: **end loop**

Because $f_{i,j}$ is a linear function, $|\frac{\partial f_{i,j}}{\partial x_d}| = |v_{i,j,d}|$ where $v_{i,j,d}$ is the d -th component of $\mathbf{v}_{i,j}$. By normalizing the sensitivities of $f_{i,j}$ with $\|\mathbf{v}_{i,j}\|_\infty$ (the infinity norm of $\mathbf{v}_{i,j}$), we can define the *importance* of the d -th variable to $f_{i,j}$ as

$$s_{i,j,d} = \frac{|v_{i,j,d}|}{\|\mathbf{v}_{i,j}\|_\infty} = \frac{|v_{i,j,d}|}{\max_k |v_{i,j,k}|}. \quad (3)$$

We argue that if a variable has low importances to all the discriminant functions then it can be identified as task-irrelevant variable. On the other hand, a variable is identified as task-relevant variable if it has unignorable importance to any $f_{i,j}$. Therefore, we define the *importance* of the d -th variable to the classification as

$$c_d = \max_{i,j \neq i} s_{i,j,d}. \quad (4)$$

c_d is the maximum of importances of the d -th variable across all discriminant functions.

Consequently, task-irrelevant variables, each of which has low importance (below a threshold) to classification, can be discarded in the unsupervised training of auto-encoder. In order to facilitate computation, we use a variable mask α to represent the binary task-relevances of input variables. α is defined as

$$\alpha = (\mathbf{1}(c_d \geq c_{th}))_{d=1}^M, \quad (5)$$

where c_{th} is a importance threshold and $\mathbf{1}(\cdot)$ is the indicator function so that it takes 1 if the condition in the brackets is true and 0 otherwise. Mask components corresponding to task-irrelevant variables will take 0.

In practice, since there might be cross-correlation between variables in input variable set, it is not easy to find out all task-irrelevant variables through training a MLR with full input variable set. A iterative method can be employed to find out task-irrelevant variables gradually, and in each iteration a new pre-classifier MLR is trained to dropping a few variables from input variable set.

Algorithm 1 is called Importance-based Variable Selection (IVS), and describes how to find task-irrelevant variables in detail. In line 1, variable mask α is initialized

by assigning 1 to each component, which means all input variables will be used in the first MLR training. For each iteration, a new pre-classifier MLR is trained with masked training data in line 3, where \odot means component-wise multiplication or Haddamard product. In order to prevent the MLR training from overfitting, model selection can be done by using a validation dataset to early stop the training in line 3. The variable mask α is updated in line 7 based on the well-trained MLR. This iterative procedure will stop under conditions such as exceeding maximum iterations, no more task-irrelevant variables found, no better classification performance obtained on validation set and so on. Once the variable mask α is obtained, task-irrelevant variables indicated by α will be dropped in the following unsupervised training for auto-encoders.

Algorithm 1 can be employed for each higher layer of stacked auto-encoders, therefore complex task-irrelevant information not eliminated on low level can be removed gradually on higher layers. From a model selection point of view, eliminating task-irrelevant variables can reduce complexity of networks therefore obtain better performance.

4 Experiments and Results

In our experiments, we combined the proposed method with DAE Vincent et al. (2010), which is called DAE-IVS, and compared performances produced by stacked DAE-IVS (SDAE-IVS), stacked DAE (SDAE) Vincent et al. (2010), PGBM Sohn et al. (2013), and aNN Wang et al. (2014). The baseline was SDAE trained with standard training strategy described in Vincent et al. (2010). We tested different depth of SDAE and SDAE-IVS from 1 layer to 3 layers. Each auto-encoder had 1000 hidden units and used tied weights. Both encoder and decoder used sigmoid function and took cross-entropy loss as reconstruction error. The outputs of feature extractors learned by a layer were used as input variables of upper layer. Multinomial Logistic Regression layer was added on both top of SDAE and SDAE-IVS to perform supervised fine-tuning. All training processes used stochastic gradient descent for parameter learning.

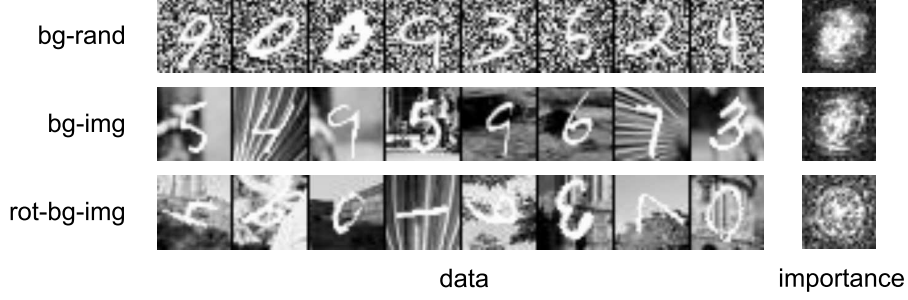


Figure 1: Inputs examples(left) and visualization of importances of variables (right). The importances are obtained from MLRs trained with full variable sets. The light of point denotes the importance of corresponding variable to classification. White stands for 1 and black for 0.

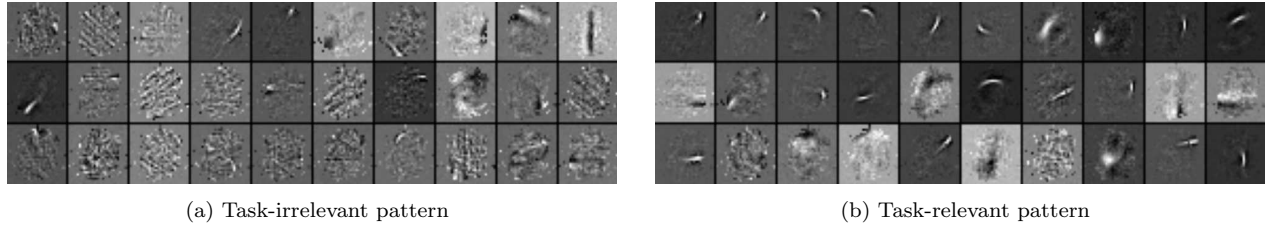


Figure 2: Visualization of learned patterns on the first hidden layer. Importance threshold is 0.3. The data set is bg-img.

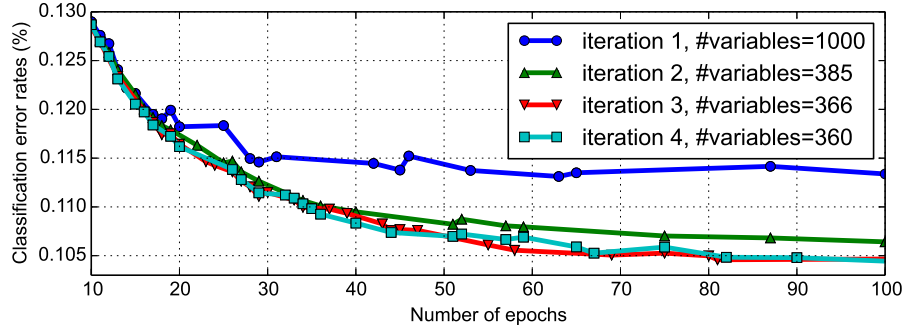


Figure 3: Test classification error rates (in %) on bg-rand at different training epochs of MLRs. Each MLR is the pre-classifier in a iteration of Algorithm 1 that performs upon the first hidden layer of SDAE-IVS. The importance threshold is 0.3, and the Gaussian noise standard deviation in SDAE-IVS is 0.2.

Our datasets were several variants of MNIST dataset for recognizing images of handwritten digits Larochelle et al. (2007), including MNIST with random background(bg-rand) or with image background (bg-img) and the combination of rotated digits with image background (rot-bg-img). Each dataset was split into three subsets: a training set (10000 examples) for pre-training and fine-tuning the parameters, a validation set (2000 examples) for model selection and a testing set (50000 examples) on which the classification performance were reported. The hyper-parameters were chosen on the validation set, including the learning rate for pre-classification in Algorithm 1 (candidate set [0.01, 0.02, 0.05, 0.1]), the importance threshold (candidate set [0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5]), learning rate for pre-training

and fine-tuning (candidate set [0.01, 0.05, 0.1, 0.2]), Gaussian noise standard deviation in SDAE and SDAE-IVS (candidate set [0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4]), and pre-training epochs (candidate set [60, 120, 180, 240, 300]). The loop in Algorithm 1 was stopped when no more task-irrelevant variables found and no better classification performance obtained on validation set.

4.1 Effect of importance-based variable selection

During Applying algorithm1, we can compute the importance value of each variable on each dataset(equation (2), (3) and (4)). Examples and visualization of the importance of variable for each dataset are shown in Figure 1. In importance image, the lighter and the darker areas correspond to the

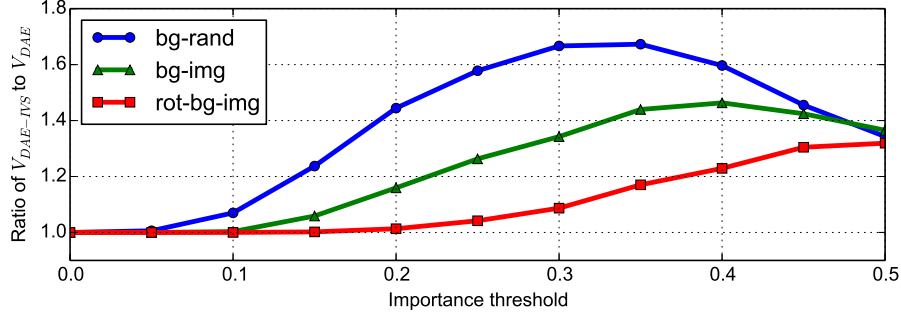
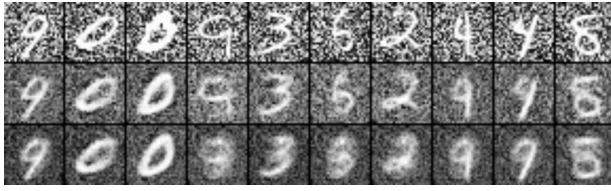
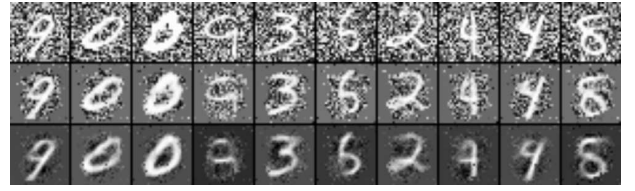


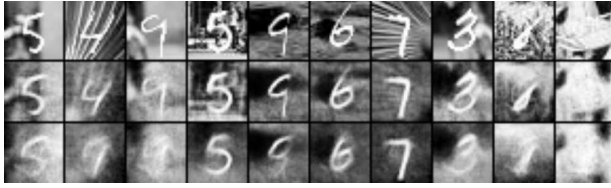
Figure 4: Ratio of $V_{DAE-IVS}$ to V_{DAE} (see the text for explanation) against importance threshold on first hidden layer.



(a) SDAE on bg-rand



(b) SDAE-IVS on bg-rand



(c) SDAE on bg-img



(d) SDAE-IVS on bg-img

Figure 5: Reconstructions produced by SDAE and SDAE-IVS. The first row is origin data, the second and the third rows are produced with 1 and 2 stacked layers respectively.

Table 1 Test classification error rates (in %) produced by different depth of SDAE and SDAE-IVS. A 95% confidence interval for each result is also given. Best performer is in bold.

	Methods	bg-rand	bg-img	rot-bg-img
1 Layer	SDAE-1	11.85 \pm 0.28	14.82 \pm 0.31	46.24 \pm 0.44
	SDAE-IVS-1	9.92 \pm 0.26	13.58 \pm 0.30	44.72 \pm 0.43
2 Layers	SDAE-2	10.00 \pm 0.26	13.41 \pm 0.30	41.05 \pm 0.43
	SDAE-IVS-2	7.58 \pm 0.23	12.31 \pm 0.28	39.35 \pm 0.43
3 Layers	SDAE-3	9.50 \pm 0.26	13.06 \pm 0.29	39.88 \pm 0.43
	SDAE-IVS-3	7.05 \pm 0.23	11.21 \pm 0.27	37.53 \pm 0.43

variables with higher importance and lower importance respectively. It can be seen that variables with high importance concentrate on the central area where digits mainly occupy. Different from other datasets, the high-importance area of rot-bg-img looks like a disc because of the rotation of digits. We take the variables with importance value higher than a threshold as the task-

relevant variable, otherwise as the task-irrelevant ones (equation (5)). Visualizations of the task-irrelevant and the task-relevant patterns (weights of feature extractors learned by the first layer of SDAE-IVS) are showed in Figure 2, where the threshold is set by experience. Although there exists misclassification, most task-irrelevant patterns describe the background image, and

most task-relevant patterns describe the foreground digit. The algorithm 1 is an iterate process, and task-irrelevant variables are dropped in each iteration. This can indirectly improve the signal to noise ratio related to classification. Figure 3 shows that not only the number of variables is decreased (to 36%), but also the classification performance of the pre-classifier is improved. Similar results can also be found on different layers of SDAE-IVS trained on different datasets.

Let $V_{DAE-IVS}$ and V_{DAE} be the number of task-relevant feature extractor learned by the first layer of SDAE-IVS and SDAE respectively. We used the ratio of $V_{DAE-IVS}$ to V_{DAE} to measure the effectiveness improvement of learning useful feature extractors. As shown in Figure 4, all these curves are above 1, which means that DAE-IVS could learn more task-relevant feature extractor than DAE. With the importance threshold increasing, many actual task-relevant feature extractors were also dropped, thus these curves get lower. However, they are still above 1. Note that we do not use the curves to optimize the threshold of importance, we just concentrate on illustrating that feature selection is beneficial to training AE. For showing the effect of feature selection, we try to reconstruct the lower-layer data through the decoders of AEs by using just the task-relevant features in higher-layer. By observing the reconstruction result, we can see whether our algorithm effectively eliminates the task-irrelevant variables and preserves the task-relevant ones. In Figure 5, we show the reconstructions of raw data produced by SDEA and SDAE-IVS with different depth on different datasets. The reconstructions are clearer after dropping task-irrelevant variables, and the background information is significantly suppressed.

4.2 Classification performance comparison

In Table 1, we show the test classification error rate of produced by SDAE and SDAE-IVS with different depth. It can be seen that in each depth the performance produced by SDAE-IVS significantly outperforms the performance of SDAE. These results suggest that our method can effectively help auto-encoders learn more and better task-relevant feature extractor so as to get better task performance.

5 Conclusion

Auto-encoders attempt to capture as much as possible of information in the input data and have to expend part of its capacity to learn task-irrelevant information if there exists. More importantly, task-irrelevant information may lead the eventual classification to overfitting resulting in bad performance.

The proposed method is a simple and effective variable selection method to deal with this problem. Through several rounds of variable selection, the remaining input variables are fed into an auto-encoder

to learn feature extractors. Because this method is employed for each layer of stacked auto-encoders, it not only eliminates task-irrelevant information, but also prunes the deep network in a certain degree so as to efficiently control the model complexity to obtain better performance. Experimental results show that the method can efficiently drop task-irrelevant variables and helps the auto-encoders learn more and better feature extractor. It helps SDAE achieve significant improvements on classification performances.

In the future, we will explore some variable selection method that deduced from non-linear classification models, expecting to help stacked auto-encoders get better performance.

Acknowledgment

This work was partially supported by the Doctoral Startup Foundation of China Three Gorges University (Grant No. KJ2013B064), Natural Science Foundation of Hubei (Grant Nos. 2015CFB336) and National Natural Science Foundation of China (Grant No. 61502274).

References

- Admuth, S. and Chile, R. (2015). Neuro-fuzzy-based hybrid controller for stable temperature of liquid in heat exchanger. *International Journal of Computational Science and Engineering*, 10(1):220 – 230.
- Arisoy, E., Sethy, A., Ramabhadran, B., and Chen, S. (2015). Bidirectional recurrent neural network language models for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5421–5425. IEEE.
- Bengio, Y. (2013). Deep learning of representations: Looking forward. In Dedi, A.-H., MartÅn-Vide, C., Mitkov, R., and Truthe, B., editors, *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 1–37. Springer Berlin Heidelberg.
- Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828.
- Bengio, Y., Lamblin, P., Larochelle, H., Popovici, D., and Montreal, U. (2007). Greedy layer-wise training of deep networks. In *In NIPS*, pages 153–160.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.
- Boureau, Y.-l., Cun, Y. L., et al. (2008). Sparse feature learning for deep belief networks. In *Advances in*

- neural information processing systems, pages 1185–1192.
- Chen, M., Weinberger, K. Q., Sha, F., and Bengio, Y. (2014). Marginalized denoising auto-encoders for nonlinear representations. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1476–1484.
- Chen, M., Xu, Z., Weinberger, K., and Sha, F. (2012). Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683*.
- Ciresan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE.
- Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.
- Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.
- Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480. ACM.
- Lee, H., Ekanadham, C., and Ng, A. Y. (2008). Sparse deep belief net model for visual area v2. In *Advances in neural information processing systems*, pages 873–880.
- Liu, N., Zheng, F., and Xia, K. (2011). An intelligent oil reservoir identification approach by deploying quantum levenberg-marquardt neural network and rough set. *International Journal of Computational Science and Engineering*, 6(1-2):76 – 85.
- Ng, A. (2011). Sparse autoencoder. *CS294A Lecture notes*, 72.
- Rifai, S., Dauphin, Y. N., Vincent, P., Bengio, Y., and Muller, X. (2011a). The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, pages 2294–2302.
- Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., and Glorot, X. (2011b). Higher order contractive auto-encoder. In *Machine Learning and Knowledge Discovery in Databases*, pages 645–660. Springer.
- Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011c). Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 833–840.
- Sohn, K., Zhou, G., Lee, C., and Lee, H. (2013). Learning and selecting features jointly with point-wise gated boltzmann machines. In *International Conference on Machine Learning*, pages 217–225.
- Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.
- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408.
- Wang, Q., Zhang, J., Song, S., and Zhang, Z. (2014). Attentional neural network: Feature selection using cognitive feedback. *ArXiv e-prints*.